H2020-EINFRA-2017

EINFRA-21-2017 - Platform-driven e-infrastructure innovation

DARE [777413] "Delivering Agile Research Excellence on European e-Infrastructures"

D6.1 Requirements and Test Cases I

Project Reference No	777413 — DARE — H2020-EINFRA-2017 / EINFRA-21-2017
Deliverable	D6.1 Requirements and Test Cases I
Work package	WP6: EPOS Use Case
Tasks involved	T6.1 Requirements Elicitation and Prioritisation
Type	R: Document, report
Dissemination Level	PU = Public
Due Date	30/06/2018
Submission Date	20/07/2018 The extension was in agreement with the PO
Status	Draft
Editor(s)	Andreas Rietbrock (KIT), Federica Magnoni (INGV), Emanuele Casarotti (INGV)
Contributor(s)	Alessandro Spinuso (KNMI), André Gemund (FRAUNHOFER)
Reviewer(s)	André Gemund (FRAUNHOFER)
Document description	This deliverable will report on the identification, assessment and prioritisation of requirements based on specific Test Cases for

	different seismological scenarios. It will take into account the needs in view of the current and envisioned European RIs. These requirements will be adjusted in later phases by taking into account the related DARE Tools and Services.

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v1	23/05/2018	Initial Structure	A. Rietbrock (KIT)
v2	06/06/2018	Test cases and available components	F. Magnoni (INGV)
v3	22/06/2018	First Version	A. Rietbrock (KIT)
v4	03/07/2018	Second Version	F. Magnoni (INGV)
v5	04/07/2018	Lineage/Provenance	A. Spinuso (KNMI)
v6	06/07/2018	Figures, Captions, Annex	F. Magnoni (INGV)
v7	09/07/2018	Third Version	E. Casarotti, F. Magnoni (INGV)
v8	19/07/2018	Final version	A. Rietbrock (KIT)

Executive Summary

This report presents the EPOS Seismological Use Case and the principal test cases that constitute its structure. The specific workflows underlying the proposed test cases and their common requirements are described with a particular focus on the test case about Rapid Assessment (RA) of seismic ground motion that represents our priority. To conclude, the currently available tools and services, as well as remaining open issues are introduced.

Table of Contents

Introduction 7
Purpose and Scope 7
Approach and relationship with other Work Packages and Deliverables 7
Methodology and Structure of the Deliverable 7

EPOS Use Case Summary 8
Motivations 8
Current scientific workflows for different seismological scenarios 8
Generic Uses Case needs: The Rapid Assessment (RA) as an example 11

Components, interfaces, tools: identifying missing parts 13
Current status: available infrastructures, e-infrastructures, interfaces, components 13
Authentication/Authorization Systems 15
Lineage/Provenance 16

Summary of Requirements 18
Open Questions 21
Missing Parts and Current Limitations 21

Conclusions 21

References 22

ANNEX 23

List of Figures

Figure 1: Workflow of Rapid Assessment (RA) test case
Figure 2: Schematic representation of the VERCE Simulation and Analysis Platform
Figure 3: Trace of data dependencies of a synthetic seismogram's image
Figure 4: View of the interactions between different computational seismology workflows
Figure 5: Data formats required by the three main test cases of the EPOS Use Case
Figure 6: Data sources to be accessed by the three main test cases of the EPOS Use Case
Figure 7: Metadata required by the three main test cases of the EPOS Use Case
Figure 8: Computational and storage requirements for the three main test cases of the EPOS Use Case
Figure A1: Workflow of the test case for point-like Seismic Source (SS) analysis with 1D wavespeed model
Figure A2: Workflow of the test case for point-like SS analysis with 3D wavespeed model
Figure A3: Workflow of the test case for finite-fault SS analysis with 1D wavespeed model
Figure A4: Workflow of the test case for finite-fault SS analysis with 3D wavespeed model
Figure A5: Workflow of Ensemble Simulation (ES) test case

List of Terms and Abbreviations

Abbreviation	Definition
RA	Rapid Assessment
SS	Seismic Sources
ES	Ensemble Simulations
GCMT	Global Centroid Moment Tensor
TDMT	Time Domain Moment Tensor

1 Introduction

1.1 Purpose and Scope

The objective of this deliverable is to identify and highlight the requirements of the EPOS generic Use Case with respect to the DARE platform. The Use Case focuses on general user needs in computational seismology building on top of developed standards and accepted interfaces in the seismological researcher and practitioner community. This will be achieved by compiling a list of requirements for the Use Case, which covers a large number of applications and workflows by the user community. The list of requirements will be used to design a proper architecture and Application Programming Interface (API) as part of the DARE platform to ensure all needs of the Use Case are covered.

1.2 Approach and relationship with other Work Packages and Deliverables

The deliverable first describes the Use Case in detail and then summarises the needs of the user community incorporating the wider research landscape in computational Earth Sciences. It is closely linked to WP2 that will design the architecture, specifically deliverable D2.1 (DARE Architecture and Technical Positioning). It is also linked to WP3 by supporting the definition of User Stories.

1.3 Methodology and Structure of the Deliverable

The structure of this deliverable is as follows. First, the generic Seismological Use Case will be summarized, followed by a description of the user needs using a detailed description of several workflows. Requirements will be extracted from those workflows and subsequently be further detailed. Components and interfaces required from an architectural point of view will be discussed afterwards.

2 EPOS Use Case Summary

This section will summarize the EPOS Use Case at a high level, along with some information about the underlying motivation and more generic aspects.

2.1 Motivation

Due to the availability of ever more seismic data and powerful synthetic simulation tools, seismologists are facing the challenge of how to analyse large amounts of data effectively and in a reliable and repeatable way. These needs become even more urgent after large earthquakes, as there is the necessity to provide rapidly reliable shaking estimates for emergency response purposes. The theoretical foundations are well established, data availability is ever increasing and the computational needs of the problem can now be accommodated by HPC resources. However, the interplay between data, computation and analysis has to be newly organized in a transparent and reliable fashion to tackle these kind of problems which will be at the centre of the EPOS Use Case in DARE.

2.2 Current scientific workflows for different seismological scenarios

Based on the work in the first few months of the project, the main test cases that compose the general EPOS Use Case have been identified delineating the underlying workflows and the principal requirements. As deeper analyses and improved understanding will develop through co-design and co-development, these initial requirements will be helpful but not over-constraining.
In the framework of EPOS Use Case within DARE, seismologists are primarily interested in:

- designing and implementing methods for Rapid Assessment (**RA**) of strong ground motion after large earthquakes also in the context of emergency response;
- the rapid characterisation of Seismic Sources (**SS**) to evaluate the impact on earthquake's wave propagation and support decision-makers in localised hazard assessments;
- on-demand Ensemble Simulations (**ES**) which are required for statistical analyses of the ground motion parameters and their uncertainties exploring the variability of the input models.

In view of these tasks, there is a strong demand for robust provenance-driven tools to organise, explore and reuse the results, with flexible management of metadata for detailed and ad-hoc validation of methods. To address these requirement, DARE should provide a holistic system that will facilitate comparative studies and will complement the rapid response to societal demands with trustworthy evidence and advice. Moreover, we can benefit from the strong experience matured during the development of the VERCE portal [Atkinson *et al*., 2015] in the framework of the VERCE and EPOS-IP projects.
Major details on the three main test cases are described in what follows and a summary of the requirements is provided in Section 4.

The **RA** of strong ground motion is considered the primary objective since most of the needed components and tools are implemented on the VERCE platform and therefore the focus can be put on integration and extension of capability of the newly deployed DARE. The aim of this first test case is to quickly analyse earthquakes and produce rapid on-demand estimates of ground motion parameters such as peak values of velocity or acceleration of the ground motion or intensity of ground shaking. Output products like waveform propagation snapshots and especially maps of ground motion parameters are fundamental for a visual representation of the earthquake. They are also useful in the

framework of emergency response, and can be compared with maps constructed based on recorded ground motion data, so-called Shakemaps [e.g. Michelini et al., 2008].
The specific steps and requirements in this case include:

1. Selecting the models to describe the region where the seismic wavefield is simulated geometrically and physically. This can be achieved by choosing a model from a library of available models or by uploading customised models. This is already implemented in the VERCE platform and it is planned to extend the available library in the course of EPOS-IP.
2. Selecting the seismic source parameters that describe the earthquake to be simulated. This can be achieved by collecting information from national and international archives (e.g. GCMT, TDMT by INGV) or uploading customised models. Both point-like seismic sources and extended fault descriptions are possible. This is already available in the VERCE platform except for the case of finite seismic sources whose usage still needs to be implemented.
3. Managing the numerical simulation software. In general, the seismological use case can use the code SPECFEM3D, already implemented in the VERCE platform as described in Section 3.1, useful both for global and local/regional seismic waveform simulations.
4. Accessing the suitable computing resources on-demand, to produce the simulated output data as quickly as possible. These data are both numerous small (tens of KB) files in ascii format (eventually converted into mseed format) for the seismograms and a smaller number of bigger (MB to GB) files in binary format for the visualization outputs. Again, this is already implemented in the VERCE platform but actual on-demand requests should be incorporated.
5. Rapid transfer of the input/output data between different co-working execution environments and storage systems, now including also Cloud resources.
6. Organisation and exploration of the runs and results based on their metadata and provenance information, for easy discovery and combination of the outputs from simulations with different inputs. This includes management tools that allow to summarise the ground motion features, combining outputs from multiple runs, while so far in the VERCE portal only one-to-one comparisons between synthetics and data are allowed.
7. Gathering of corresponding observed data from national and international available archives (e.g. EIDA, INGV Shakemaps); these data can be seismograms in seed format (as already managed by the VERCE platform) but now also binary files containing information on the strong ground motion parameters extrapolated from the analysis of observed data like Shakemaps [Michelini *et al.*, 2008].
8. Managing the tools requested in Section 3.1 for the comparison and combination of synthetic outputs on earthquake strong ground motion and the corresponding information based on observed data. The flux of input and output information exchanged during these procedures is usually codified by ascii, xml or json/geojson files.
9. Handling the storage requirements. For a complete RA experiment, the volume of data to be stored can reach a maximum of tens of terabytes per user in the production phase.
10. Handling the computing demand. For a complete RA experiment, the computational resource requirements can reach a maximum of tens of millions of CPU hours per user in the production phase.

The **SS** analysis aims at characterizing the parameters of earthquake sources like the earthquake location, magnitude and rupture mechanism represented by the so-called moment tensor. In case of modelling the earthquake as an extended fault, the parameters include the values and direction of the displacement that occurred on this fault. In SS analysis the simulated synthetic waveforms for an initial model of the seismic source are compared to the observed data in order to invert for improved values

of the source parameters (minimizing this misfit) and to estimate the associated uncertainties. These parameters and uncertainties characterize the seismic sources and are fundamental for further hazard assessment analyses.

The RA and SS cases have the steps and requirements described at points 1-7 above in common. Then the impact caused by the seismic source on the ground motion parameters are analysed. In the framework of seismic source analysis, we plan to distinguish four different test cases depending on the model chosen for the earthquake source and for the wavespeed structure (see also the Annex):

- study of point-like seismic sources using 1D wavespeed models;
- study of point-like seismic sources using 3D wavespeed models;
- study of seismic sources modelled as slip on a fault with finite dimensions using 1D wavespeed models;
- study of seismic sources modelled as slip on a fault with finite dimensions using 3D wavespeed models.

With respect to RA, these cases involve additional simulations or access to pre-calculated basis-function libraries required by the inversion procedures implemented in the software packages of Section 3.1. For example, point source inversions in 3D require 6 to 9 additional simulations for each earthquake obtained by perturbing the source parameters one-at-a-time [Liu et al., 2004]. The other three cases require calculation of seismic wavefields for unitary input sources, i.e. pre-calculated so-called Green's functions, forming the basis functions that are combined by the inversion procedures to get updated source solutions (e.g. [Dreger *et al.*, 2005]). In this sense, as described in point 6 of RA, multiple simulations for the same earthquake should be easily linked based on metadata and provenance, in order to combine the input for the inversions. Among these four test cases in which SS is articulated, we consider the study of point-like sources with 3D wavespeed models a priority, especially because part of the workflow has been already experienced in VERCE and we agreed on a main tool for inversion with a straightforward implementation (Section 3.1).

The format of input/output data for SS task is, as in the RA case, ascii/xml/json for the summary files of the analysis softwares and ascii/seed files for the seismograms. Analyses for the SS tasks also involve the code FLEXWIN/pyflex, described in Section 3.1, for the selection of waveform time windows suitable for inversion procedures, whose usage is already managed by the VERCE platform. Finally, the storage and computational requirements described at points 9 and 10 for task RA above are also valid for the SS task.

The **ES** task has the scope of statistically characterizing the ground motion parameters and their uncertainties, analysing ensembles of models constructed by the variability of the input parameters. Thus, it shares the requirements at points 1-7 described for RA, but in this case we are more focused on exploring the variability of the source model parameters. The earthquake source can be modelled as points or finite faults, and their impact on ground motion assessment, highlighting the strong connection of this test case with the other two proposed test cases. At step 2 of the RA test case, rather than requiring the selection of a single source model, it should be allowed to perform a grid search on ranges of values of the source parameters, implying that for each value of the range a new simulation should be carried out, while the other input parameters stay fixed. Thus, a major characteristic of this task is that a very large number of simulations (hundreds to thousands) or a library of pre-calculated basis Functions (e.g. Green's functions) will be required. Their outputs should be managed automatically, also implementing tools to summarize them for comparisons with observations (requirement 7) and to quickly and easily link to these results in order to use them as input of the

software for ensemble and uncertainty analyses described in Section 3.1 (requirement 8). Other specific requirements are:

- Handling a storage demand that can reach a maximum of hundreds of TB per user for a complete ES experiment in the production phase.
- Handling a computing demand that ranges from tens to hundreds of millions of CPU hours per user for a complete ES experiment in the production phase.

Metadata and provenance information will be crucial to discover preliminary results for further integration and comparison. As already anticipated, intermediate results, such as those describing the unitary functions produced in support of the SS and ES, will have to be discoverable according to methods' parameterisation and contextual results' metadata. This will allow to trigger automatic configuration of the extended simulation workflows with the desired model perturbation.

For the RA, data products such as Shakemaps obtained from the large synthetic data will require to be properly identifiable based on the ground motion parameters, in order to create the statistical sample, which will be used for comparison with the Shakemap products based on observed data and to refine the model used for the assessment. We foresee that the concrete practice and progress will improve the description of the processes and their output. In order to achieve this aim flexibility and retrieval performance of metadata and lineage management is important to guarantee that new experimental results can be managed, rapidly discovered, combined and most importantly evaluated to create new better constrained runs.

Moreover, in order to support validation and effective management of results, it should be possible to organise the workflow executions that produced the statistical samples visually, highlighting their contribution to the progress of the model refinement in a retrospective analysis, if any, or suggesting the elimination of their results, thereby freeing a substantial amount of resources.

2.3 Generic Uses Case needs: Rapid Assessment (RA) as an example

The workflow that constitutes the structure of the Rapid Assessment (RA) test case at a high level is summarized in Figure 1.

The first step consists in defining a model to describe the physical properties of the wave propagation medium, i.e. a wavespeed model, and a model for the geometry of the medium, i.e. a grid, usually called mesh, that discretizes the volume. As already in the VERCE portal, users should be allowed to either select these input elements from an available library (for example internal to the platform), or to upload their own files.

Another step in the collection of input data is the selection of a model to describe the earthquake source, either point-like source models or finite fault models. These models could for example be chosen from public archives of seismic source solutions (e.g. GCMT, TDMT solutions from INGV), while still allowing users to upload their own files. In case of finite fault description, a library internal to the portal could be created offering possible source solutions for given earthquakes selectable by users.

In order to facilitate the selection of the above described inputs, an extreme flexibility of the platform is required since the input elements should be gathered from different data sources (see Section 4), different data formats should be managed (see Section 4), and users should also be able to customize all of the model parameters.

After input selection, users should choose a simulation code and the parameters for the seismic waveform simulations. Thanks to the VERCE project, we have extensive experience with the code SPECFEM3D (see Section 3.1), both in the version for global seismic simulations and for local/regional ones. Nevertheless, this step also requires a strong flexibility, since the code has many functionalities, is continuously updated, and all its parameters should be easily adjustable by the users. Moreover, since numerous other codes exist and continue to be developed for waveform simulation, a big aim is that the code selection becomes as much interchangeable as possible in order to make the platform usable by many research groups that use different procedures.

At this point of the workflow, the RA test case requires the calculation of ground motion parameters (as peak ground values of displacement, velocity and acceleration) from the simulated wavefields. We thus need to implement tools to extract these parameters based on a specific procedure. As for the simulation code, different research groups use different methodologies and tools for the analyses. Thus, the design of this step should allow the possibility of implementing multiple procedures and, in case, include some interface where users can customize the analysis routines depending on the experiment carried out.

Then, maps of ground motion parameters based on observed data for the same earthquake that has been simulated can be gathered from public archives, as the Shakemaps from INGV[1]. In this case the main requirements are the management of multiple data formats and data sources (see Section 4).

Finally, we need to implement tools to compare the Shakemap based on observed data and those constructed using the synthetic simulations. Moreover, it could be useful to take into account procedures that allow the integration of synthetic information with observed data in order to produce more realistic ground motion maps that include all the complexities of the Earth structure and of the seismic source modelled through the simulations. The step of comparison and misfit calculation should be very general, since it is present in the workflows of all the other test cases as well (see the Annex), and should be able to accommodate multiple procedures and approaches used by different research groups, directly customizable within the platform.

To conclude, the output products of the workflow, like integrated shakemaps and summaries of calculated ground motion parameters, should be stored along with their metadata and complete provenance information. The choice of the needed metadata and provenance to be extracted and stored will be based on an initial design of the analysis procedure, but should be flexible enough in order to include additional or modified values that become useful when updated or different methodologies are implemented.

At least the first part of the described workflow, until the waveform simulation step, is already implemented in the VERCE platform, which should ease its implementation on the new DARE platform. Moreover, many of the steps are shared with the other test cases presented in Section 2.2 (see figures in the Annex). This is highlighted by the coloured dots in Figure 1 where green represents the SS test case and blue represents the ES case. The RA test case thus exemplifies a generic high-level workflow that will be implemented in the DARE platform and represents the main requirements that should be taken into account for the EPOS Use Case. It is fundamental that all the components of this workflow will be as abstract and adaptive as possible in order to be executed on different environments (Cloud, HPC, local resources) depending on the availability and user needs, and to be reusable in the workflows of the other proposed test cases. This will facilitate the implementation of the EPOS Use Case in its entirety achieving most of the envisioned goals. Moreover, this will make the created platform an extremely valuable product that can be used for specific scopes, but that is also adaptable to fully customized scientific experiments and continuously evolving analysis procedures. This enlarges the

[1] http://shakemap.rm.ingv.it/shake/archive/

audience of potentially interested users and would guarantee a large future success and exploitation of the platform after the end of the DARE project.

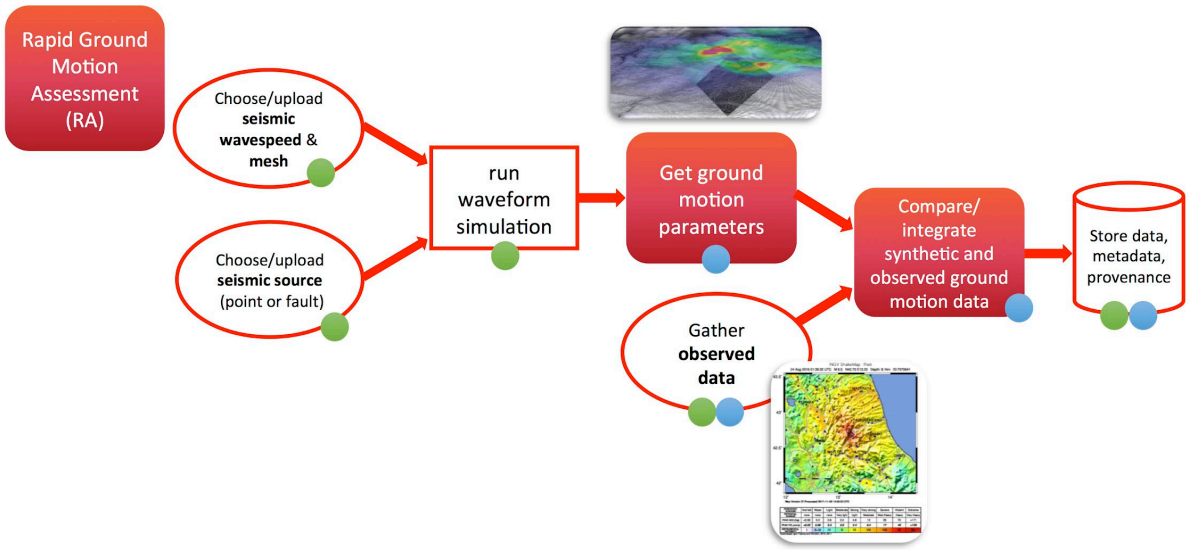

Figure 1: Steps of the workflow for the Rapid Assessment test case. The coloured dots associated with each element indicate the steps that are in common with the other proposed test cases (compare with the figures in Annex): green is for the Seismic Source (SS) analysis test cases, blue for the Ensemble Simulation (ES) test case.

3 Components, interfaces, tools: identifying missing parts

3.1 Current status: available infrastructures, e-infrastructures, interfaces, components

A main component for the computational seismology use case is the eScience scientific **portal**[2] developed during the European FP7 project **VERCE** [Atkinson *et al.*, 2015] and refined and rolled out in the European FP7 project EPOS-IP. This platform has been developed to allow both expert and less expert users to quickly perform simulations of the seismic wavefield generated by an earthquake and to easily manage post-processing and analyses of the output data. The portal functionalities are carried out through four principal workflows:

1. the **Simulation Workflow** allows users to select the simulation region with a corresponding seismic wavespeed model (by choosing from an implemented library or by uploading bespoke models), select the earthquake to be simulated, the seismic stations, and to finally launch the simulation run;
2. the **Download Workflow** permits to query seismological European archives for raw recorded seismograms corresponding to the simulated waveforms;
3. the **Processing Workflow** allows to apply typical seismological steps to both observed and simulated traces in order to prepare them for comparison;
4. the **Misfit Workflow** offers different procedures to calculate the misfit values between recorded and simulated seismograms, fundamental to study e.g. the model behaviour or to approach waveform inversion; this workflow has been strongly improved and updated in the current release of the portal.

[2] https://portal.verce.eu/home

The main software implemented in the VERCE portal for waveform simulations is **SPECFEM3D** [Peter *et al.* 2011], both its version for local/regional simulations and the one for regional/global scale. This is a Fortran 95 code largely tested worldwide and scalable on a huge number of cores and also on GPU resources. Moreover, for the misfit calculation, two other codes are already implemented in the portal: the first code is **pyflex** (L. Krisher[3]), a python port of the Fortran 95 code FLEXWIN (Maggi *et al.* 2009). It selects time windows on the seismograms, where it calculates cross-correlation misfit parameters between observed and synthetic traces. The second code is the python code developed by Kristekova *et al.* [2006, 2009] that calculates time-frequency misfit criteria on full seismic waveforms.

The intensive numerical calculations of the VERCE portal are performed exploiting HPC resources of EGI and PRACE computing centres, and recent updates tested the readiness of different cloud providers of the EGI Federated Cloud[4] to support the EGI VO through which the portal is running.
The processing steps of workflows run by the VERCE platform are controlled behind the scenes by another fundamental component, the cross-platform processing framework **dispel4py**[5]. This is a python library specifically designed to describe abstract workflows for distributed data-intensive applications and to allow their execution in a large variety of parallel environments. This component thus represents the base of the VERCE platform workflows, orchestrating the management of input and output data, the connections and relationships between the different workflows, down to the definition of the fundamental pipelines that constitute the single processing steps within the platform. This level of abstraction and granularity guarantees the strong flexibility of the portal allowing for easy customization of the procedures by the users and for a continuous update of the portal functionalities in order to support the evolving requirements from field researchers. It is moreover clear that dispel4py can cover a key role both for the seismological use case and for the climatological one within DARE.

In this framework, python and especially its package **ObsPy**[6] are other essential components for computational seismology applications. ObsPy is a widely adopted python framework for processing seismological data; it provides parsers for common file formats, clients to access data centres and fundamental seismological signal processing routines, which allow the manipulation of seismological time series. The processing elements managed by dispel4py are all written in python using the ready-to-use functions specifically designed for the needs of seismological researchers by ObsPy.

Finally, the VERCE portal also counts on the functioning of external services that allow for gathering the input data used within the portal. For example, the web service of Federation of Digital Seismographic Network (**FDSN**) is an option implemented in the portal to collect the parameters of the earthquakes and stations to be used in the simulations. A second example is that in the Download Workflow, the observed seismograms for waveform comparisons can be searched and downloaded through **Orfeus/EIDA** nodes.

All the above described components are already in place and operational under the EPOS umbrella and are considered fundamental for the development of the seismological use case within DARE. In addition, other required components are some of the seismological software packages already used by

[3] http://krischer.github.io/pyflex/
[4] https://www.egi.eu/federation/egi-federated-cloud/
[5] https://github.com/dispel4py/dispel4py
[6] http://doi.org/10.5281/zenodo.165135

the seismic community to address the main tasks planned in the EPOS Use case (WP6), some examples are:

a) pycmt3d[7] is the code that we plan to use for point-like moment tensor source inversions with 3D wavespeed models;
b) other codes or libraries for seismic source inversions modelled as point sources e.g. pyTDMT[8], tensorflow[9], instaseis[10];
c) codes for finite source inversions e.g. Dreger *et al.* [2005];
d) code for shakemap calculation (e.g. the one implemented at INGV[11]);
e) some state-of-the-art library for machine learning analysis (ES) like tensorflow[8] and scikit-learn[12].

3.2 Authentication/Authorization Systems

As the test cases makes extensive use of distributed services provided by multiple e-Infrastructures, Authentication and Authorization is a key challenge to overcome during the implementation. Researchers that instrument workflows using the User Interface on the Science Gateway need to be authenticated not only the Gateway itself, but at the services of the other infrastructure involved in the execution of the workflow as well. For example, data retrieval from data centres through their provided web services might require authentication, the backend web services of the Gateway require authentication, compute resources that carry out simulations or data processing require authentication, and the iRODS infrastructure for permanent result storage requires authentication as well. As an additional complexity, some of the services are controlled through different protocols. For example, at least during the time of the aforementioned VERCE project, job submission and data transfer to HPC systems typically required the use of the Grid Security Infrastructure (GSI)[13] family of protocols. File transfers were carried out using the GridFTP protocol, and job submissions through the Globus GRAM[14] interface, both of which required the use of X.509 certificates[15]. Unfortunately, there is no one Authentication system that supports all involved components. To address this complexity, the current infrastructure makes use of the following components:

a) A LDAP server for authentication data used by the Science Gateway and iRODS (can be considered a community Identity Provider in current terms). This was also used by resource providers to retrieve account information when corresponding agreement existed.
b) A VOMS server as attribute authority to certify group membership and Roles (used mainly for EGI resources)
c) A MyProxy server to store proxy certificates for GridFTP and GRAM compatibility
d) A Java WebStart application to upload a certificate to a MyProxy Server
e) The gUSE framework credential components

[7] https://github.com/wjlei1990/pycmt3d
[8] http://webservices.rm.ingv.it/pyTDMT/
[9] https://www.tensorflow.org
[10] http://instaseis.net
[11] http://shakemap.rm.ingv.it
[12] http://scikit-learn.org
[13] https://web.archive.org/web/20010527095836/http://www.globus.org/Security/overview.html
[14] http://toolkit.globus.org/toolkit/docs/6.0/gram5/index.html
[15] https://tools.ietf.org/html/rfc5280

Considering the current developments through projects like EOSC-hub[16] and AARC[17], we hope that we can reduce the complexity of the Authentication and Authorization mechanisms by relying on new federated OpenID-Connect[18] based proxy services and connected credential services such as RCauth[19] for protocols relying on X.509 certificates. First prototypes and evaluations have already been carried out and led to promising results.

3.3 Lineage/Provenance

In the aforementioned project VERCE, several technical challenges were tackled when integrating provenance extraction and management mechanisms. In some cases, these were required by high security standards imposed by the computational services, or by metadata and identifier schemas that had to be consistently represented within the lineage. Moreover, the processing services, offered by the platforms where the provenance system had to be integrated, were of quite different scale in terms of size and computational needs. These went from simple and rapid data processing operations on small portions of a dataset, to large simulations and postprocessing tasks involving many files of different formats that required long runs of multiple interconnected workflows. In this context, provenance-driven tools enable rapid exploration of results and of the relationships between data, which accelerates understanding, method improvements and semi-automatic configuration of interdependent workflows and workspaces, as shown in Figure 2.

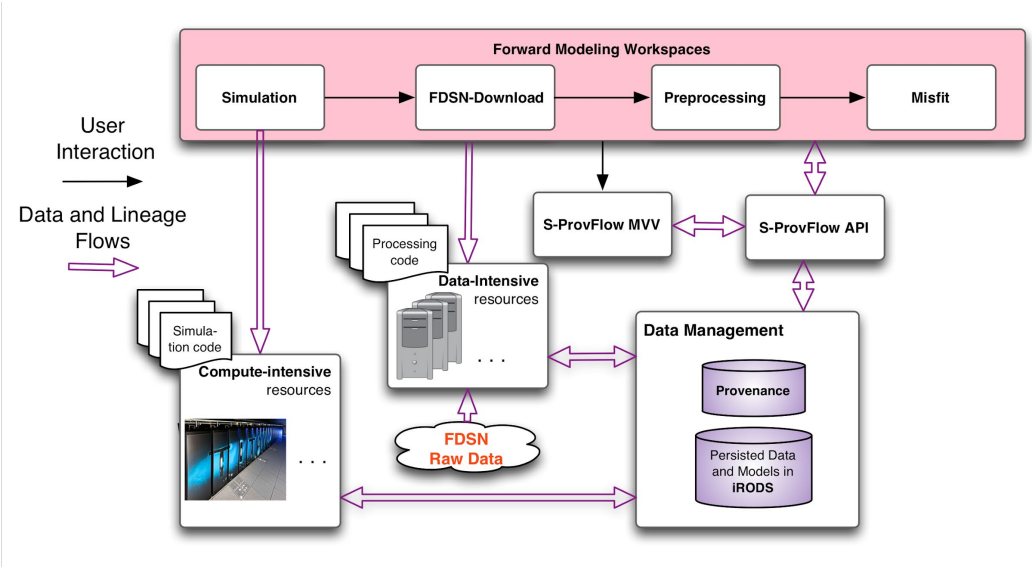

Figure 2: VERCE Simulation and Analysis Platform: Schematic representation depicting the *Users* workspaces and their interaction with the *System* components through exchange of data and lineage. All the workspaces, from the simulation to the misfit are controlled by the users that access interactively the provenance services to discover and combine the data produced by the previous phases (or by previous runs of the same phase), and that will be involved in the configuration and the input of the next workflow. All workspaces can be operated independently.

[16] https://eosc-hub.eu/
[17] https://aarc-project.eu/
[18] http://openid.net/connect/
[19] https://rcauth.eu/

The workspaces are implemented through a combination of batch processes and workflows developed with the dispel4py data-intensive tool. They all generate lineage documents that are ingested by *S-ProvFlow* and linked in a coherent trace as shown in Figure 3. This is achieved by populating the S-PROV provenance model, which is developed on top of existing standards such as PROV, ProvONE[20] and the vocabulary offered by SEIS-PROV[21] to represent metadata about computational artefacts in seismology. Such model, in combination with the provenance framework adopted by the workflow system, demonstrated to be flexible enough to serve different use cases. For instance, the provenance data exposed through the *S-ProvFlow* API is used during each of these phases to gather relevant information to setup the inputs and the parameters associated with the workflows and to perform activities such as validation, results management and comprehensive analysis on data reuse and exploitation of the resources. This was achieved by developing different classes of visualisation tools adopting also advanced visual analytics techniques like shown in Figure 4.

In DARE we will develop methods where provenance mechanisms are required to produce metadata-rich traces for tailored data-products generated by the three main use cases (RA, SS, ES). We have demands for robust provenance-driven tools to organise, explore and reuse the results derived by the ensembles and the rapid assessment analysis, with flexible management of metadata for detailed and ad-hoc validation of their methods. The holistic system already experimented in VERCE and further developed within DARE should facilitate comparative studies and should complement the rapid response to societal demands with trustworthy evidence and advice.

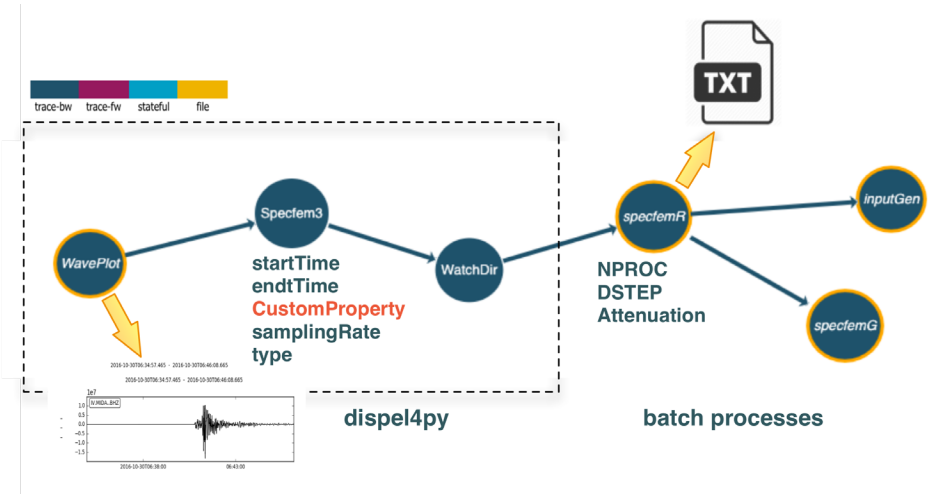

Figure 3: Trace of data dependencies (backwards navigation) of a synthetic seismogram's image. The section in the box is performed by a data-intensive workflow, while the rest is implemented via batch processes. All generate lineage documents that are received by S-ProvFlow asynchronously at runtime and combined in a coherent visualisation. Each intermediate data is described by a set of metadata terms belonging to agreed vocabularies or defined by the user and specific to the experiment. Some data dependencies link to actual data files (yellow circles) that can be accessed interactively for detailed analysis and reuse.

[20] https://purl.dataone.org/provone-v1-dev

[21] http://seismicdata.github.io/SEIS-PROV/

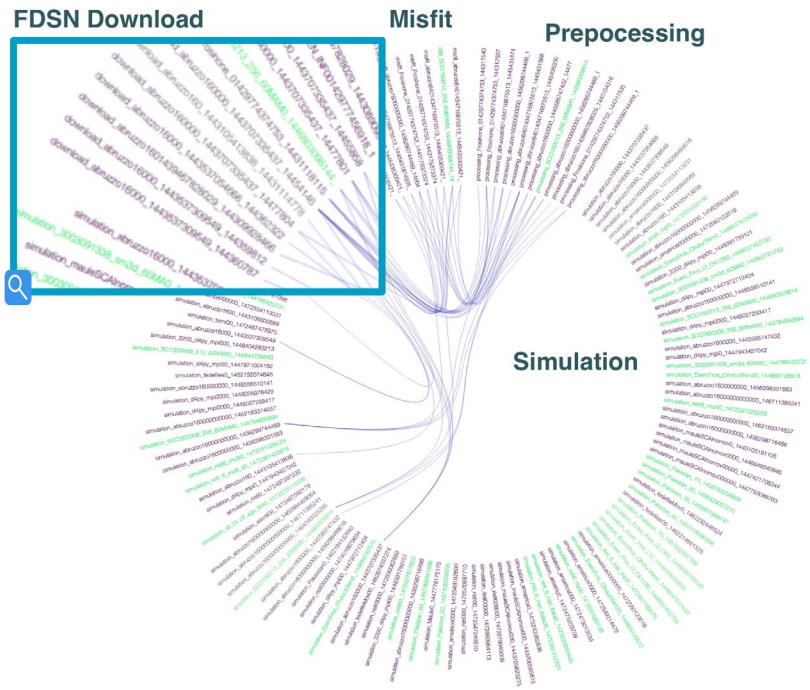

Figure 4: View of the interactions between different computational seismology workflows grouped by their type: *Simulation*, *FDSN-Download*, *Processing* and *Misfit*. The colour of the vertices indicate different users, while the magnifier shows a download run whose results have been reused by many preprocessing tasks, suggesting the presence of a good dataset or the target of a particular investigation. The view is obtained through S-ProvFlow that offers this as an interactive tool on top of the S-PROV model.

4 Summary of Requirements

Analysing the interconnections and overlapping steps between the test cases described in Section 2.2, we have identified common requirements that will be the base for the work of the architectural task force (WP2) and for the construction of the user stories (task T3.1 OF WP3).

- All the test cases require the combination of numerous outputs from multiple workflows, thus all the outputs should be described by their detailed metadata and provenance to allow their exploration, reuse and combination for complex analyses (see Section 3.3).
- Since the proposed use cases have been designed with many overlapping steps, all the workflows that constitute their structure should be built with a high modularity in order to be as general as possible and to be applicable to different processing. This increases the platform flexibility and the possibility of adapting it to evolving approaches.
- The three test cases described in Section 2.2 also require that all the involved execution environments (HPC, Clouds and institutional resources) should be quickly and easily linked to each other in order to reduce and optimize the time required for analyses and transfers of data.
- Another requirement is the possibility of handling different **data formats** for input and output products. Although this is again a general requirement, the three use cases have specific formats to be handled and details are reported in Figure 5. The figure lists the formats required for the involved input and output files for each test case, and the specific products managed by every case are highlighted with colours. As the main goal of EPOS Use Case is studying the variation of groundmotion parameters caused by earthquake source variability, It is evident that overlaps

between the requirements of the different test cases exist and especially that ES in general combines the needs of RA and SS test cases.

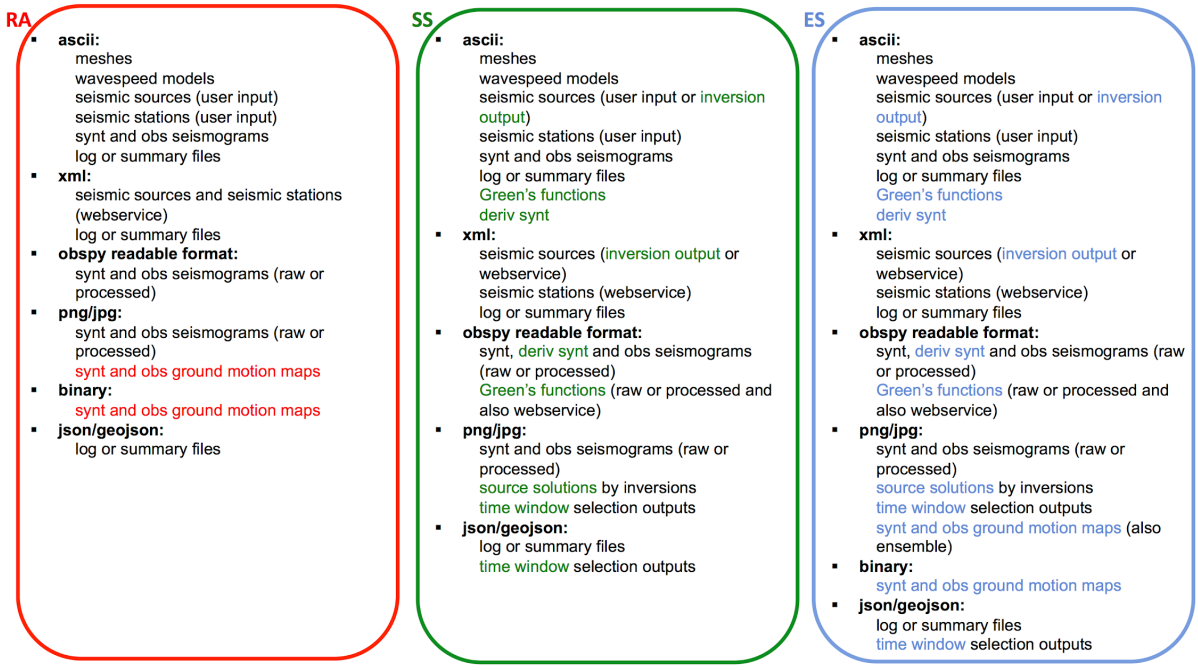

Figure 5: Data formats required for the input and output files of the three main test cases of the EPOS Use Case. Specific products managed by every case are highlighted with corresponding colours.

- To gather the input products of the test cases the exploitation of multiple data sources is required. The specific requirements slightly vary between the three use cases, further details are provided in Figure 6. In the RA case, it is required to access public repositories of data for the source solutions, stations, waveforms and ground motion parameters. However, users should also be allowed to upload their own inputs for customized experiments (see Sections 2.2 and 2.3). The same is valid for the SS test case that moreover requires to access public or private repositories of Green's functions and source geometry models. The ES test case again combines all the previous requirements.

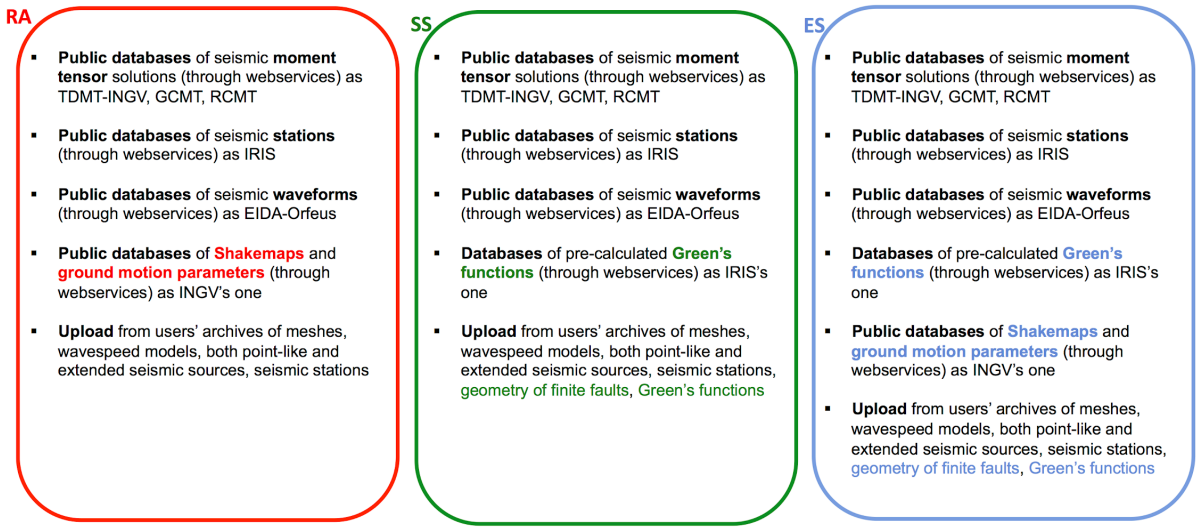

Figure 6: Data sources to be accessed in order to gather the input files of the three main test cases of the EPOS Use Case. Specific products managed by every case are highlighted with corresponding colours.

- Following the strong requirement of carefully describing all the inputs and outputs of the test cases with detailed metadata and provenance in order to make them searchable and reusable, in Figure 7 we report a list of the main metadata that should be attributed and stored for the products of the use cases. Thus, RA specifically involves metadata for the ground motion maps, while SS involves metadata to describe the Green's functions and the inverted source models. As previously, ES combines all these requirements.

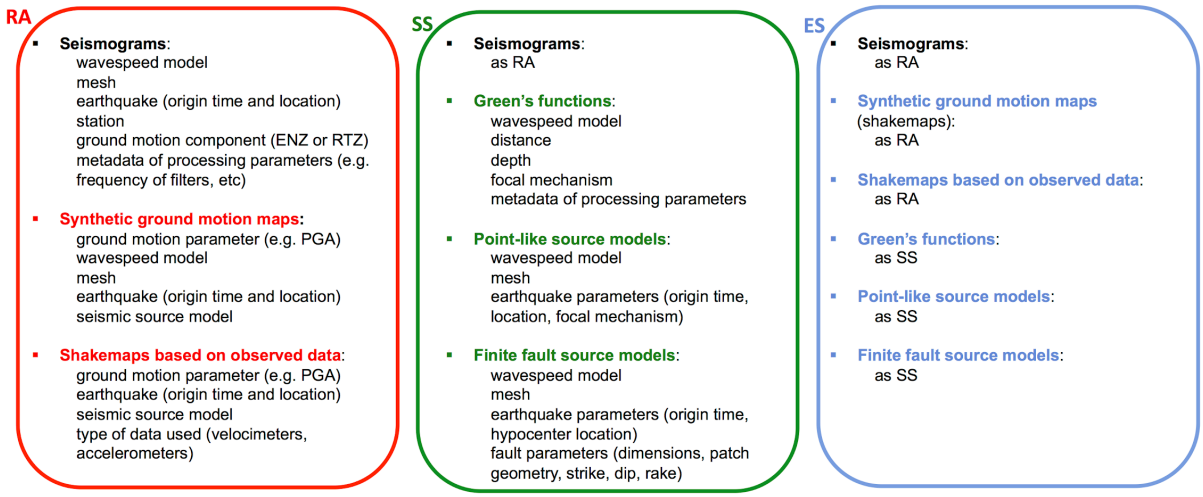

Figure 7: Metadata to be attributed and stored to the input and output files of the three main test cases of the EPOS Use Case. Colours highlight the specific products to be handled by every test case.

- The last important requirement is the storage and computing demand. Figure 8 gives an upper limit of the resources that would be required for complete experiments in the production stage, i.e. when the platform will be finally deployed and usable (as anticipated in Section 2.2). However, there are possibilities to reduce these demands, for example by using pre-calculated basis functions that can be recombined instead of performing new simulations every time. Moreover, in the development phase both computing and storage requirements are drastically lower (~100s CPUhs and few TBs per user).

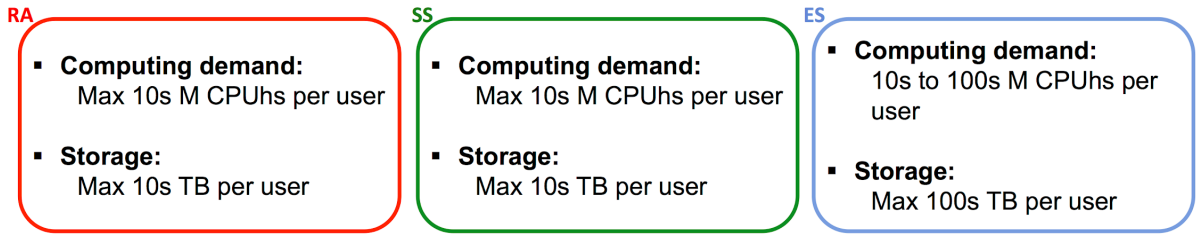

Figure 8: Computational and storage requirements for complete experiments in the context of the three main test cases of the EPOS Use Case, considering a production phase.

The requirements and workflows underlying the main test cases of the EPOS Use Case are also detailed in the tables compiled to support WP3 in the definition of the user stories and of the common requirements between the principal DARE Use Cases from WP6 and WP7. In general, we expect that

the described requirements and workflows will be defined with more details during the next steps of the project.

4.1 Open Questions

In the next months, a few points need to be further addressed and are at this stage considered open questions:

a) What are the main resources (computational environments, computing hours, storage amount) available for the EPOS Use Case? In general, the use case requires a large amount of computational resources, even larger in a productive setting. The availability of such resources could be assured in different ways, thanks to the interaction of European High-Performance Computing centres and PRACE.

b) Do we want to reuse the components already implemented in the VERCE portal or start the implementation of the EPOS Use Case from scratch? As seen in Section 3, the VERCE portal provides several components already in a productive stage, an unbeatable advantage for the EPOS use case. Nevertheless, the request of more flexibility and additional components could be better achieved implementing from scratch with new technology.

c) In case we want to start from the VERCE platform, what components, tools and services will be reused? In this case, the choice of the useful components should be based on the specific workflows of the proposed test cases and taking into account the overlapping between their steps (see Sections 2.2, 2.3 and Annex).

d) Do we want to introduce a metadata standard for the seismological use case not used so far for the VERCE portal? The VERCE portal handles provenance preservation in realtime during the different steps of simulations and processing, resulting in a metadata-rich environment. The scheme follows the W3C PROV scheme. We should modify such scheme in order to assume the metadata and keyword specific mapping developed for seismology inside the WP8 of EPOS-IP.

4.2 Missing Parts and Current Limitations

The principal limitation at this stage is the lack of a high level of flexibility especially for developing new techniques to compare synthetic and observed data. One solution could be to allow users to upload their own processing elements (e.g. jupyter-notebooks) which need to be embedded in an API allowing the users to access the simulation outputs and observed data easily. Additionally, we should explore the viability of simple visualisations using image file formats or even more sophisticated interfaces like WorldWind (https://worldwind.arc.nasa.gov/) to display results. Users should then also be able to add provenance information to their computation which will be explorable later on. As HPC security very often is a limitation, the computational part should be abstracted in reusable workflows which hopefully will be accepted by HPC centres as already shown in the VERCE project.

5 Conclusions

To summarise the conclusions of the present deliverable, regarding the test cases, the priority will be given to the RA test case and then to the analyses of point-like SS with 3D wavespeed models. In this way we focus in the beginning mainly on the integration and customisation of the workflow elements into the DARE platform and then developing new tools taking full advantage of the DARE technology.

The 1D analyses of the earthquake seismic sources are nevertheless fundamental, since they do not require strong computational demands and the large availability of 1D models will enlarge the audience of interested users. 1D tests are also valuable and quick benchmarks for more complex analyses.

Regarding the requirements, higher priority is on the strong exploitation of metadata and provenance in order to combine numerous multiple workflows, the flexibility and abstraction of the workflow pipelines, and quick and easy transfers and analyses of data also for emergency responses, together with the specific requirements for the RA test case about metadata, data sources and formats.

Finally, we so far envisage two testbeds with different degree of data availability for two seismic regions of the world with high societal and economic impact:

1) Italy, mainly central Italy, for which we have a huge amount of high quality data and experienced procedures;
2) Greece.

6 References

[Atkinson *et al.* 2015] M.P. Atkinson, M. Carpené, E. Casarotti, S. Claus, R. Filgueira, A. Frank, M. Galea, T. Garth, A. Gemünd, H. Igel, I. Klampanos, A. Krause, L. Krischer, S. H. Leong, F. Magnoni, J. Matser, A. Michelini, A. Rietbrock, H. Schwichtenberg, A. Spinuso, and J.-P. Vilotte, *VERCE delivers a productive e-Science environment for seismology research*, in Proc. IEEE eScience 2015.

[Dreger *et al.* 2005] D. S. Dreger, L. Gee, P. Lombard, M. H. Murray, & B. Romanowicz, 2005. *Rapid finite-source analysis and near-fault strong ground motions: Application to the 2003 Mw 6.5 San Simeon and 2004 Mw 6.0 Parkfield earthquakes*. Seismol. Res. Lett., 76(1), 40–48.

[Liu *et al.* 2004]

[Maggi *et al.* 2009] A. Maggi, C. Tape, M. Chen, D. Chao, & J. Tromp, 2009. *An automated time window selection algorithm for seismic tomography, Geophys*. J. Int., 178, 257–281.

[Michelini *et al.* 2008] A. Michelini, L. Faenza, V. Lauciani & L. Malagnini 2008. *ShakeMap implementation in Italy*. Seismological Research Letters, 79(5), 688–697.

[Kristeková *et al.* 2006] M. Kristeková, J. Kristek, P. Moczo, & S. M. Day 2006. *Misfit Criteria for Quantitative Comparison of Seismograms*. Bulletin of the Seismological Society of America, 96(5), 1836–1850. http://doi.org/10.1785/012006

[Kristeková *et al.* 2009] M. Kristeková, J. Kristek & P. Moczo, 2009. *Time-frequency misfit and goodness-of-fit criteria for quantitative comparison of time signals*. Geophysical Journal International, 178(2), 813–825. http://doi.org/10.1111/j.1365- 246X.2009.04177.x

[Peter *et al.* 2011] D. Peter, D. Komatitsch, Y. Luo, R. Martin, N. Le Goff, E. Casarotti, P. Le Loher, F. Magnoni, Q. Liu, C. Blitz, T. Nissen-Meyer, P. Basini & J. Tromp, 2011. *Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes*, Geophys. J. Int., 186, 721–739.

7 ANNEX

The following figures present the workflows underlying the Seismic Source (SS) test cases and the Ensemble Simulation (ES) test case.

In particular, for SS we distinguish between four different test cases (Figures A1-A4) depending on the source representation and wavespeed model. The earthquake source can be modelled as a point-like source or as a slip on a fault with finite dimensions. The seismic wavespeed can be modelled in 1D or a 3D structure (see Section 2.2). Note that for the study of finite seismic sources with 3D wavespeed models (FIgure A4), the general workflow is the same as in the 1D case (Figure A3). The main differences are, beyond the considered structure model, the Green's functions that should be constructed, that now have different dependencies, and the code used for the inversion that should take into account the 3D complexity of the model.

Comparing all the workflows, it is evident that many steps are shared between them, which is highlighted in the figure by dots whose colours refer to every specific test case. This overlapping between the workflow steps would facilitate the implementation and reusability of the developed components.

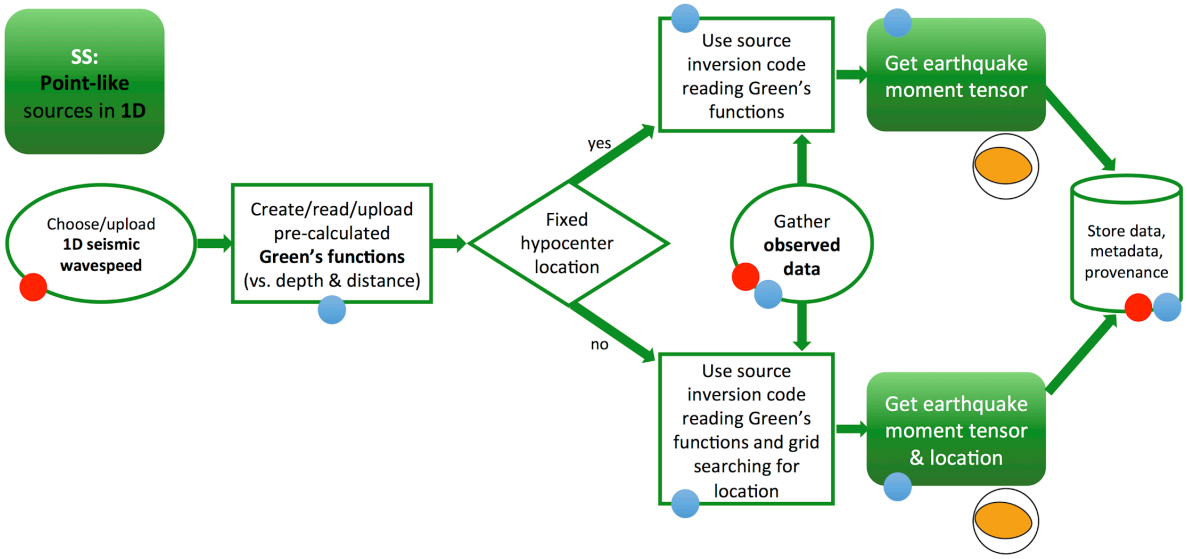

Figure A1: Workflow's steps of the test case for point-like Seismic Source (SS) analysis with 1D wavespeed model. The coloured dots associated with each element indicate the steps that are in common with the other proposed test cases (compare with Figure 1 and A5): red is for the Rapid Assessment (RA) test case, blue for the Ensemble Simulation (ES) test case.

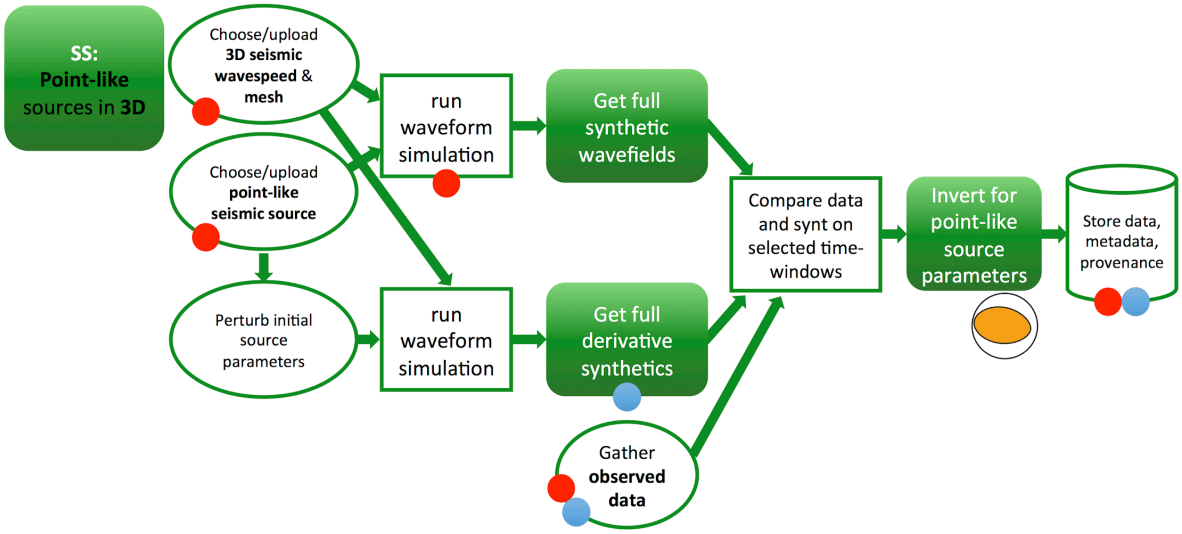

Figure A2: Workflow's steps of the test case for point-like Seismic Source (SS) analysis with 3D wavespeed model. The coloured dots associated with each element indicate the steps that are in common with the other proposed test cases (compare with Figure 1 and A5): red is for the Rapid Assessment (RA) test case, blue for the Ensemble Simulation (ES) test case.

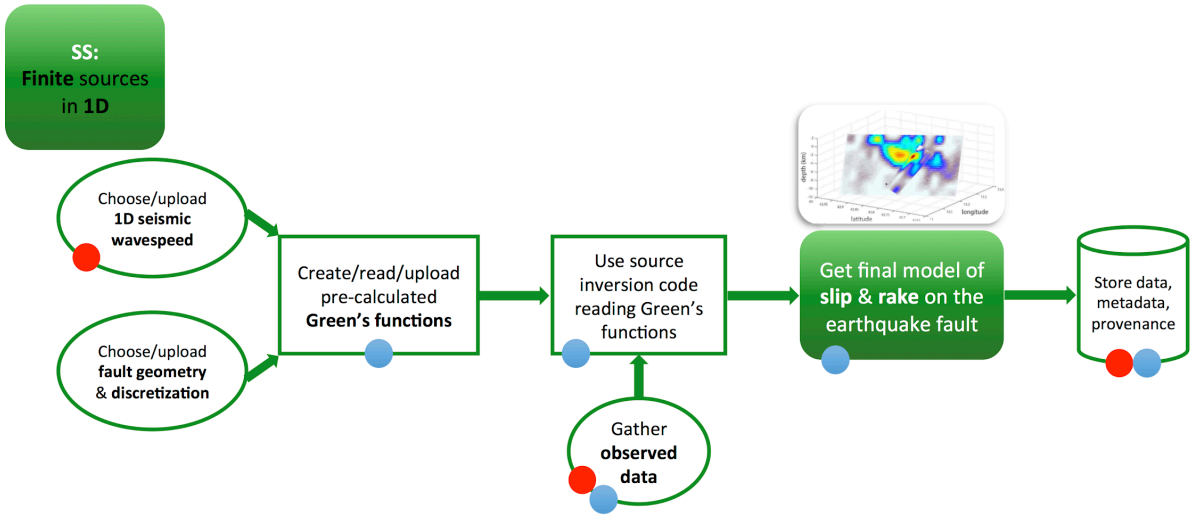

Figure A3: Workflow's steps of the test case for finite fault Seismic Source (SS) analysis with 1D wavespeed model. The coloured dots associated with each element indicate the steps that are in common with the other proposed test cases (compare with Figure 1 and A5): red is for the Rapid Assessment (RA) test case, blue for the Ensemble Simulation (ES) test case.

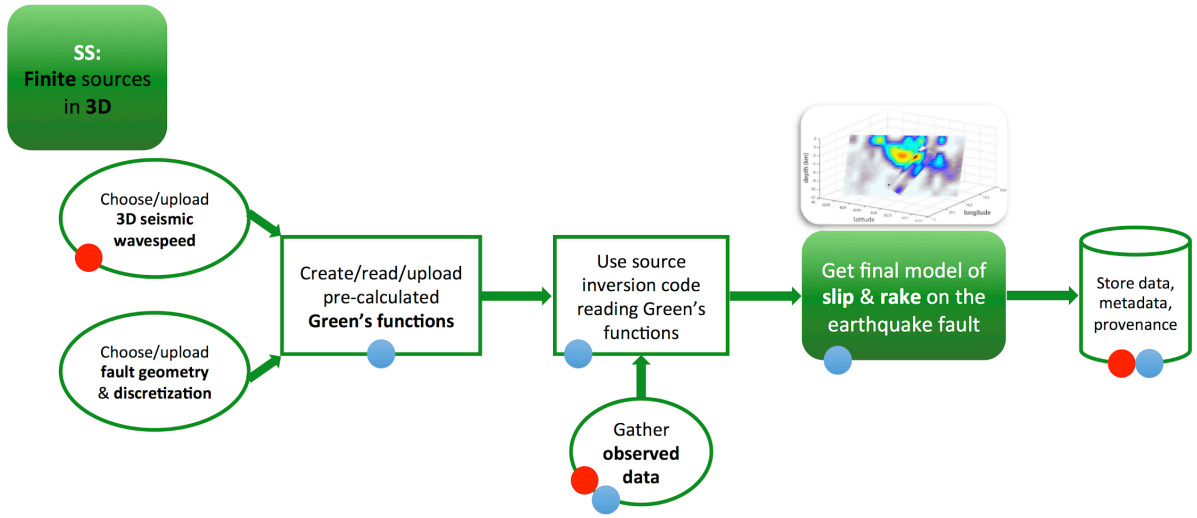

Figure A4: Workflow's steps of the test case for finite fault Seismic Source (SS) analysis with 3D wavespeed model. The coloured dots associated with each element indicate the steps that are in common with the other proposed test cases (compare with Figure 1 and A5): red is for the Rapid Assessment (RA) test case, blue for the Ensemble Simulation (ES) test case.

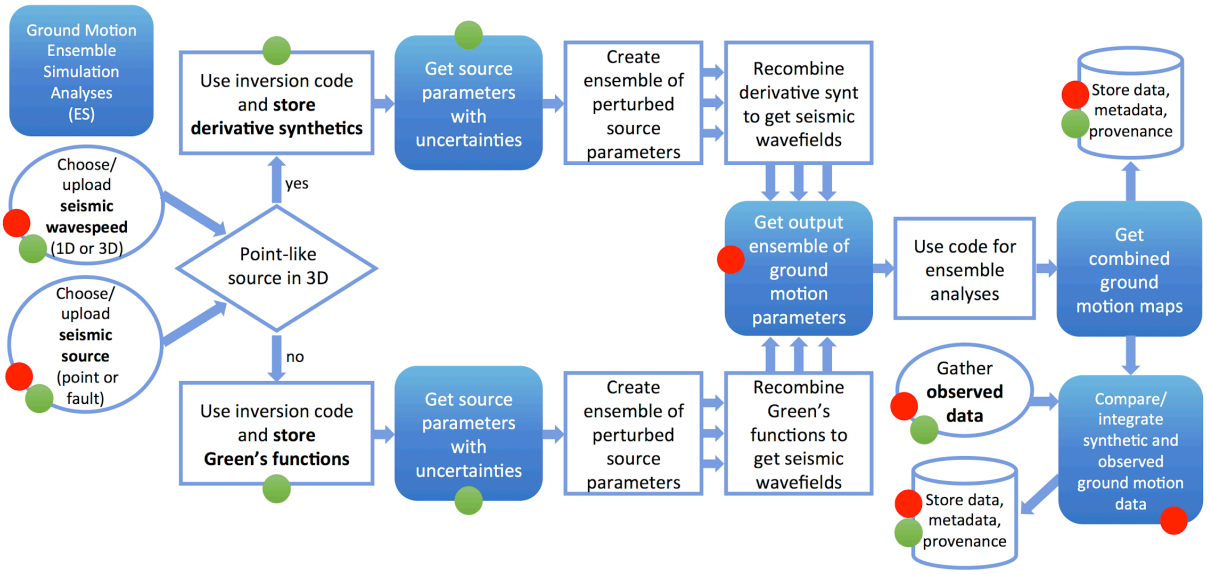

Figure A5: Steps of the workflow for the Ensemble Simulation (ES) test case. The coloured dots associated with each element indicate the steps that are in common with the other proposed test cases (compare with Figures 1 and A1-A4): green is for the Seismic Source (SS) analysis test cases, red is for the Rapid Assessment (RA) test case.